

Report ANOVA_Posthoc_R_Slides

Malik Haroon

Advanced Statistical Programming (ACCA7022)

```

> cat("=====\\n")
=====
> cat("STEP 1: LOADING AND EXPLORING THE GAPMINDER DATASET\\n")
STEP 1: LOADING AND EXPLORING THE GAPMINDER DATASET
> cat("=====\\n\\n")
=====

>
> # Load the dataset
> data("gapminder")
> cat("Dataset loaded successfully!\\n")
Dataset loaded successfully!
> cat("Dimensions:", dim(gapminder)[1], "rows and", dim(gapminder)[2], "columns\\n\\n")
Dimensions: 1704 rows and 6 columns

>
> # Display first few rows
> cat("First few rows of the dataset:\\n")
First few rows of the dataset:
> print(head(gapminder))
# A tibble: 6 × 6
  country    continent year lifeExp  pop gdpPercap
<fct>      <fct>    <int> <dbl> <int> <dbl>
1 Afghanistan Asia    1952  28.8 8425333  779.
2 Afghanistan Asia    1957  30.3 9240934  821.
3 Afghanistan Asia    1962  32.0 10267083  853.
4 Afghanistan Asia    1967  34.0 11537966  836.
5 Afghanistan Asia    1972  36.1 13079460  740.
6 Afghanistan Asia    1977  38.4 14880372  786.

>
> # Check structure
> cat("\\nDataset structure:\\n")

Dataset structure:
> str(gapminder)
tibble [1,704 × 6] (S3: tbl_df/tbl/data.frame)
 $ country : Factor w/ 142 levels "Afghanistan",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ continent: Factor w/ 5 levels "Africa","Americas",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ year    : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
 $ lifeExp : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop     : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372
12881816 13867957 16317921 22227415 ...

```

```

$ gdpPercap: num [1:1704] 779 821 853 836 740 ...
>
> # Summary statistics
> cat("\nSummary statistics:\n")

Summary statistics:
> print(summary(gapminder))
  country    continent    year    lifeExp
Afghanistan: 12 Africa :624 Min. :1952 Min. :23.60
Albania : 12 Americas:300 1st Qu.:1966 1st Qu.:48.20
Algeria : 12 Asia :396 Median :1980 Median :60.71
Angola : 12 Europe :360 Mean :1980 Mean :59.47
Argentina : 12 Oceania : 24 3rd Qu.:1993 3rd Qu.:70.85
Australia : 12 Max. :2007 Max. :82.60
(Other) :1632
  pop    gdpPercap
Min. :6.001e+04 Min. : 241.2
1st Qu.:2.794e+06 1st Qu.: 1202.1
Median :7.024e+06 Median : 3531.8
Mean :2.960e+07 Mean : 7215.3
3rd Qu.:1.959e+07 3rd Qu.: 9325.5
Max. :1.319e+09 Max. :113523.1

>
> # Check number of observations per continent
> cat("\nNumber of observations per continent:\n")

Number of observations per continent:
> print(table(gapminder$continent))

Africa Americas Asia Europe Oceania
 624 300 396 360 24

```

```

> #=====
> # 2. VISUALIZE GDP PER CAPITA BY CONTINENT
> #=====
>
> cat("\n===== \n")

=====
> cat("STEP 2: VISUALIZING GDP PER CAPITA BY CONTINENT\n")
STEP 2: VISUALIZING GDP PER CAPITA BY CONTINENT
> cat("===== \n\n")

```

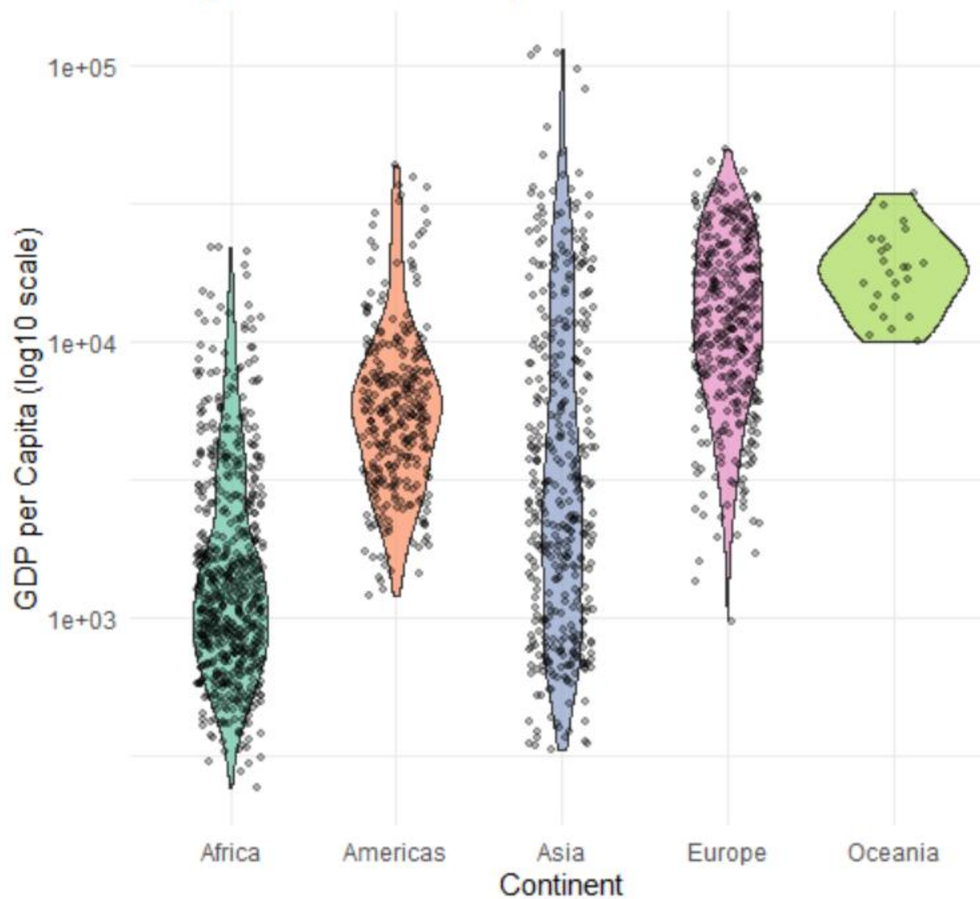
```

=====
>
> # Create boxplot
> p1 <- ggplot(gapminder, aes(x = continent, y = gdpPercap, fill = continent)) +
+   geom_boxplot() +
+   scale_y_log10() + # Log scale because GDP varies widely
+   labs(title = "GDP per Capita by Continent",
+     subtitle = "Log10 scale for better visualization",
+     x = "Continent",
+     y = "GDP per Capita (log10 scale)") +
+   theme_minimal() +
+   theme(legend.position = "none") +
+   scale_fill_brewer(palette = "Set2")
>
> print(p1)
> readline(prompt = "\nPress [Enter] to see next plot...")

Press [Enter] to see next plot...
[1] ""
> # Create violin plot with jittered points
> p2 <- ggplot(gapminder, aes(x = continent, y = gdpPercap, fill = continent)) +
+   geom_violin(alpha = 0.7) +
+   geom_jitter(width = 0.2, alpha = 0.3, size = 1) +
+   scale_y_log10() +
+   labs(title = "Distribution of GDP per Capita by Continent",
+     subtitle = "Violin plot with individual data points",
+     x = "Continent",
+     y = "GDP per Capita (log10 scale)") +
+   theme_minimal() +
+   theme(legend.position = "none") +
+   scale_fill_brewer(palette = "Set2")
>
> print(p2)
> readline(prompt = "\nPress [Enter] to continue to ANOVA...")

```

Distribution of GDP per Capita by Continent
Violin plot with individual data points



```
> #=====
> # 3. CONDUCT ONE-WAY ANOVA
> #=====
>
> cat("\n===== \n")

=====
> cat("STEP 3: CONDUCTING ONE-WAY ANOVA\n")
STEP 3: CONDUCTING ONE-WAY ANOVA
> cat("===== \n\n")

=====

>
> # Perform ANOVA
> anova_model <- aov(gdpPercap ~ continent, data = gapminder)
>
```

```

> # Display ANOVA summary
> cat("ANOVA Summary:\n")
ANOVA Summary:
> print(summary(anova_model))
      Df Sum Sq Mean Sq F value Pr(>F)
continent  4 3.799e+10 9.498e+09 126.6 <2e-16 ***
Residuals 1699 1.275e+11 7.504e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> # Extract and display key statistics
> anova_summary <- summary(anova_model)[[1]]
> cat("\nKey Findings:\n")

Key Findings:
> cat("F-statistic:", round(anova_summary$`F value`[1], 2), "\n")
F-statistic: 126.57
> cat("Degrees of freedom:", anova_summary$Df[1], "and", anova_summary$Df[2], "\n")
Degrees of freedom: 4 and 1699
> cat("p-value:", format(anova_summary$`Pr(>F)`[1], scientific = TRUE), "\n")
p-value: 1.172126e-94
>
> # Interpretation
> cat("\nInterpretation:\n")

Interpretation:
> if(anova_summary$`Pr(>F)`[1] < 0.05) {
+   cat("✓ p < 0.05 → Reject null hypothesis\n")
+   cat("✓ At least one continent has a significantly different mean GDP per capita\n")
+   cat("✓ Post-hoc tests are needed to identify which continents differ\n")
+ } else {
+   cat("✗ p >= 0.05 → Fail to reject null hypothesis\n")
+   cat("✗ No significant differences found between continents\n")
+ }
✓ p < 0.05 → Reject null hypothesis
✓ At least one continent has a significantly different mean GDP per capita
✓ Post-hoc tests are needed to identify which continents differ

```

```

=====
> cat("STEP 4: CHECKING ANOVA ASSUMPTIONS\n")
STEP 4: CHECKING ANOVA ASSUMPTIONS
> cat("=====\n\n")

```

```

=====
>
> # Set up diagnostic plots
> par(mfrow = c(2,2))
> plot(anova_model)
> par(mfrow = c(1,1))
>
> cat("Diagnostic plots displayed above:\n")
Diagnostic plots displayed above:
> cat("1. Residuals vs Fitted - Checks linearity and homoscedasticity\n")
1. Residuals vs Fitted - Checks linearity and homoscedasticity
> cat("2. Normal Q-Q - Checks normality of residuals\n")
2. Normal Q-Q - Checks normality of residuals
> cat("3. Scale-Location - Checks homoscedasticity\n")
3. Scale-Location - Checks homoscedasticity
> cat("4. Residuals vs Leverage - Checks for influential points\n\n")
4. Residuals vs Leverage - Checks for influential points

>
> # Formal tests for assumptions
> cat("Formal Assumption Tests:\n")
Formal Assumption Tests:
>
> # Test for homogeneity of variances (Levene's test)
> library(car)
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

  recode

> levene_test <- leveneTest(gdpPercap ~ continent, data = gapminder)
> cat("\nLevene's Test for Homogeneity of Variances:\n")

Levene's Test for Homogeneity of Variances:
> print(levene_test)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  4 47.256 < 2.2e-16 ***
      1699
---

```

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> if(levene_test$`Pr(>F)`[1] > 0.05) {
+   cat("✓ Variances are homogeneous (p > 0.05)\n")
+ } else {
+   cat("✗ Variances are not homogeneous (p < 0.05)\n")
+   cat("  Consider using Welch's ANOVA or data transformation\n")
+ }
✗ Variances are not homogeneous (p < 0.05)
  Consider using Welch's ANOVA or data transformation
>
> # Test for normality of residuals
> shapiro_test <- shapiro.test(residuals(anova_model))
> cat("\nShapiro-Wilk Test for Normality of Residuals:\n")

Shapiro-Wilk Test for Normality of Residuals:
> print(shapiro_test)

      Shapiro-Wilk normality test

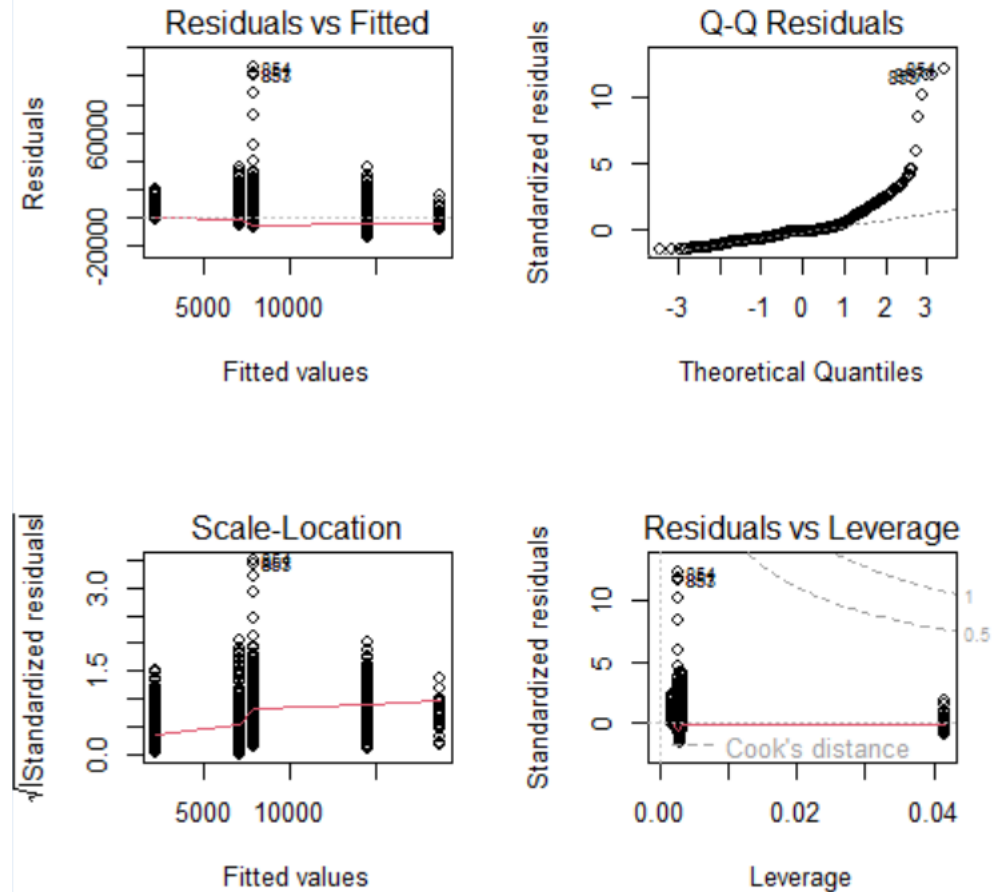
data: residuals(anova_model)
W = 0.66823, p-value < 2.2e-16

> if(shapiro_test$p.value > 0.05) {
+   cat("✓ Residuals are normally distributed (p > 0.05)\n")
+ } else {
+   cat("✗ Residuals are not normally distributed (p < 0.05)\n")
+   cat("  ANOVA is robust to moderate violations with large sample sizes\n")
+ }
✗ Residuals are not normally distributed (p < 0.05)
  ANOVA is robust to moderate violations with large sample sizes
>
> # Sample sizes per group
> cat("\nSample sizes per continent:\n")

Sample sizes per continent:
> print(table(gapminder$continent))

Africa Americas  Asia  Europe Oceania
  624    300   396   360    24
> cat("\nANOVA is robust to normality violations when sample sizes are large

```



```

=====
> cat("STEP 5: PERFORMING TUKEY HSD POST-HOC TEST\n")
STEP 5: PERFORMING TUKEY HSD POST-HOC TEST
> cat("=====\\n\\n")
=====

>
> # Perform Tukey HSD test
> tukey_result <- TukeyHSD(anova_model)
> cat("Tukey's Honest Significant Differences:\\n")
Tukey's Honest Significant Differences:
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gdpPercap ~ continent, data = gapminder)

```

```

$continent
      diff    lwr    upr    p adj
Americas-Africa 4942.3558 3280.4806 6604.231 0.0000000
Asia-Africa     5708.3958 4188.6340 7228.158 0.0000000
Europe-Africa   12275.7210 10710.1623 13841.280 0.0000000
Oceania-Africa  16427.8546 11507.4034 21348.306 0.0000000
Asia-Americas   766.0401 -1044.5150 2576.595 0.7767582
Europe-Americas 7333.3652 5484.2011 9182.529 0.0000000
Oceania-Americas 11485.4989 6467.6035 16503.394 0.0000000
Europe-Asia     6567.3251 4844.7530 8289.897 0.0000000
Oceania-Asia    10719.4588 5746.8215 15692.096 0.0000000
Oceania-Europe  4152.1337 -834.6909 9138.958 0.1539474

>
> # Extract and display results in a more readable format
> tukey_df <- as.data.frame(tukey_result$continent)
> tukey_df$comparison <- rownames(tukey_df)
> tukey_df <- tukey_df[, c("comparison", "diff", "lwr", "upr", "p adj")]
> names(tukey_df) <- c("Comparison", "Mean_Difference", "Lower_CI", "Upper_CI",
"Adjusted_p")
>
> cat("\nFormatted Tukey HSD Results:\n")

Formatted Tukey HSD Results:
> print(tukey_df)
      Comparison Mean_Difference Lower_CI Upper_CI
Americas-Africa Americas-Africa 4942.3558 3280.4806 6604.231
Asia-Africa      Asia-Africa    5708.3958 4188.6340 7228.158
Europe-Africa    Europe-Africa  12275.7210 10710.1623 13841.280
Oceania-Africa   Oceania-Africa  16427.8546 11507.4034 21348.306
Asia-Americas    Asia-Americas   766.0401 -1044.5150 2576.595
Europe-Americas  Europe-Americas  7333.3652 5484.2011 9182.529
Oceania-Americas Oceania-Americas 11485.4989 6467.6035 16503.394
Europe-Asia      Europe-Asia     6567.3251 4844.7530 8289.897
Oceania-Asia     Oceania-Asia    10719.4588 5746.8215 15692.096
Oceania-Europe   Oceania-Europe  4152.1337 -834.6909 9138.958
      Adjusted_p
Americas-Africa 2.797540e-12
Asia-Africa     2.775113e-12
Europe-Africa   2.727152e-12
Oceania-Africa  2.775780e-12
Asia-Americas   7.767582e-01
Europe-Americas 2.770228e-12

```

```

Oceania-Americas 5.165886e-09
Europe-Asia      2.773559e-12
Oceania-Asia     4.733304e-08
Oceania-Europe   1.539474e-01
>
> # Identify significant differences
> significant <- tukey_df[tukey_df$Adjusted_p < 0.05, ]
> cat("\nSignificant Differences (p < 0.05):\n")

Significant Differences (p < 0.05):
> if(nrow(significant) > 0) {
+   print(significant[, c("Comparison", "Mean_Difference", "Adjusted_p")])
+ } else {
+   cat("No significant pairwise differences found.\n")
+ }

      Comparison Mean_Difference Adjusted_p
Americas-Africa Americas-Africa   4942.356 2.797540e-12
Asia-Africa      Asia-Africa    5708.396 2.775113e-12
Europe-Africa    Europe-Africa   12275.721 2.727152e-12
Oceania-Africa   Oceania-Africa   16427.855 2.775780e-12
Europe-Americas Europe-Americas    7333.365 2.770228e-12
Oceania-Americas Oceania-Americas  11485.499 5.165886e-09
Europe-Asia      Europe-Asia     6567.325 2.773559e-12
Oceania-Asia     Oceania-Asia    10719.459 4.733304e-08
>

```

```

=====
> cat("STEP 6: VISUALIZING POST-HOC RESULTS\n")
STEP 6: VISUALIZING POST-HOC RESULTS
> cat("=====\\n\\n")
=====

>
> # Method 1: Plot Tukey HSD confidence intervals
> par(mar = c(5, 8, 4, 2)) # Increase left margin for labels
> plot(tukey_result, las = 1, col = "blue")
> title(main = "Tukey HSD Confidence Intervals\\nDifferences in Mean GDP per Capita by
Continent")
> par(mar = c(5, 4, 4, 2)) # Reset margins
>
> readline(prompt = "\\nPress [Enter] to see next visualization...")

Press [Enter] to see next visualization...

```

```
[1] ""
> # Method 2: Create a more detailed visualization with ggplot2
> # Calculate means and confidence intervals for each continent
> continent_stats <- gapminder %>%
+   group_by(continent) %>%
+   summarise(
+     mean_gdp = mean(gdpPercap),
+     se_gdp = sd(gdpPercap) / sqrt(n()),
+     lower_ci = mean_gdp - qt(0.975, df = n() - 1) * se_gdp,
+     upper_ci = mean_gdp + qt(0.975, df = n() - 1) * se_gdp,
+     n = n()
+   ) %>%
+   ungroup()
>
> # Create bar plot with confidence intervals
> p3 <- ggplot(continent_stats, aes(x = reorder(continent, mean_gdp), y = mean_gdp, fill =
continent)) +
+   geom_bar(stat = "identity", alpha = 0.7) +
+   geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci), width = 0.2) +
+   labs(title = "Mean GDP per Capita by Continent with 95% Confidence Intervals",
+     subtitle = "Bars show means, error bars show 95% CI",
+     x = "Continent",
+     y = "Mean GDP per Capita") +
+   theme_minimal() +
+   theme(legend.position = "none") +
+   scale_fill_brewer(palette = "Set2") +
+   coord_flip() # Flip coordinates for better label readability
>
> print(p3)
> readline(prompt = "\nPress [Enter] to see next visualization...")
```

Press [Enter] to see next visualization...

```
[1] ""
> # Method 3: Alternative approach - Create a heatmap of p-values
> # Extract Tukey results and create a matrix of adjusted p-values
> tukey_matrix <- as.matrix(tukey_result$continent)
> tukey_pvals <- tukey_matrix[, "p adj"]
>
> # Create a data frame for the heatmap
> comparisons <- names(tukey_pvals)
> group1 <- sapply(strsplit(comparisons, "-"), "[", 1)
> group2 <- sapply(strsplit(comparisons, "-"), "[", 2)
>
> pval_df <- data.frame(
```

```

+ Group1 = group1,
+ Group2 = group2,
+ P_Value = tukey_pvals,
+ Significant = ifelse(tukey_pvals < 0.05, "Yes", "No")
+ )
>
> # Create a heatmap-like visualization
> p4 <- ggplot(pval_df, aes(x = Group1, y = Group2, fill = P_Value)) +
+   geom_tile(color = "white") +
+   geom_text(aes(label = sprintf("%.4f", P_Value)), size = 3) +
+   scale_fill_gradient(low = "red", high = "white",
+     limits = c(0, 1),
+     name = "Adjusted\nP-Value") +
+   labs(title = "Tukey HSD Post-hoc Test Results",
+     subtitle = "Red cells indicate significant differences (p < 0.05)",
+     x = "Continent", y = "Continent") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
>
> print(p4)
> readline(prompt = "\nPress [Enter] to see next visualization...")

Press [Enter] to see next visualization...
[1] ""
> # Method 4: Create a boxplot with significance annotations
> # First, create a function to add significance stars
> add_significance <- function(p_value) {
+   ifelse(p_value < 0.001, "***",
+     ifelse(p_value < 0.01, "**",
+       ifelse(p_value < 0.05, "*", "ns")))
+ }
>
> # Add significance information to the comparisons
> pval_df$Significance_Star <- add_significance(pval_df$P_Value)
>
> # Find the maximum y-value for placing significance annotations
> y_max <- max(gapminder$gdpPercap, na.rm = TRUE) * 1.2
>
> # Create a basic boxplot
> p5 <- ggplot(gapminder, aes(x = continent, y = gdpPercap, fill = continent)) +
+   geom_boxplot(alpha = 0.7) +
+   scale_y_log10() +
+   labs(title = "GDP per Capita by Continent with Pairwise Significance",
+     subtitle = "Significance codes: *** p<0.001, ** p<0.01, * p<0.05, ns = not significant",

```

```

+   x = "Continent",
+   y = "GDP per Capita (log10 scale)" ) +
+   theme_minimal() +
+   theme(legend.position = "none") +
+   scale_fill_brewer(palette = "Set2")
>
> # Add significance annotations for key comparisons
> # You can add specific comparisons manually
> # For example, add annotation for Africa-Europe comparison
> africa_europe_p <- pval_df[pval_df$Group1 == "Africa" & pval_df$Group2 == "Europe",
"Significance_Star"]
> if(length(africa_europe_p) > 0) {
+   p5 <- p5 + annotate("text", x = 1.5, y = 50000,
+     label = paste("Africa-Europe:", africa_europe_p),
+     size = 4, color = "red")
+ }
>
> print(p5)
> readline(prompt = "\nPress [Enter] to see next visualization...")

```

Press [Enter] to see next visualization...

[1] ""

```

> # Method 5: Create a pairwise comparison matrix plot
> # This is a more sophisticated approach using the ggpubr package
> # Install if not available
> if(!require(ggpubr)) {
+   install.packages("ggpubr")
+   library(ggpubr)
+ }

```

Loading required package: ggpubr

```

>
> # Create a list of comparisons for ggpubr
> my_comparisons <- list(
+   c("Africa", "Europe"),
+   c("Africa", "Americas"),
+   c("Africa", "Asia"),
+   c("Africa", "Oceania"),
+   c("Americas", "Europe"),
+   c("Asia", "Europe"),
+   c("Oceania", "Europe")
+ )
>
> # Create boxplot with significance bars
> p6 <- ggboxplot(gapminder, x = "continent", y = "gdpPercap",

```

```

+ fill = "continent", palette = "Set2",
+ title = "GDP per Capita by Continent with Significance Tests",
+ subtitle = "Pairwise comparisons using Tukey HSD") +
+ scale_y_log10() +
+ stat_compare_means(comparisons = my_comparisons,
+ method = "t.test", # Using t-test as proxy
+ label = "p.signif") +
+ theme_minimal() +
+ theme(legend.position = "none")
>
> print(p6)

```

